**Báo Cáo Tuần( Task 2.2)**

Thời gian từ ngày 29/11/2020 – 6/12/2020

**1. Danh sách nhóm DA\_KTLT:**

* Lộc Quốc Huy – 1812369
* Nguyễn Việt Hưng – 1812499
* Trần Quang Huy – 1812427

**2. Các công việc đã thực hiện:**

Nhóm đã cải thiện source code các crawler giúp việc crawl dễ dàng hơn, đồng thời sử dụng Dataframe trong thư viện Pandas để format data theo các field và lưu trữ bộ Dataset (~10.000 item) dưới dạng csv. Trong đó công việc được phân công cụ thể:

* Tìm hiểu về thư viện Pandas và các cách format dữ liệu bằng Dataframe trong Pandas đã push báo cáo lên Github. (Trần Quang Huy)
* Crawl data theo format từ Tiki và Lazada đã push lên Github. (Trần Quang Huy)
* Crawl data theo format từ Nguyen Kim và Dienmaycholon đã push lên Github. (Nguyễn Việt Hưng)
* Crawl data theo format từ Shopee và Viết báo cáo tuần. (Lộc Quốc Huy)
* Tìm hiểu về cách làm sạch dữ liệu và đã push lên Github (Nguyễn Việt Hưng)

**3. Các khó khăn gặp phải:**

Trong quá trình làm task nhóm gặp phải các khó khăn sau:

* Source code phải thêm thủ công nhiều đường link mới crawl nhiều được. Nên nhóm đã sử dụng vòng lặp thay đổi số page của một category sản phẩm và đã push lên Github.
* Khi mở file csv dataset bằng excel thì bị lỗi font tiếng việt nên nhóm đã sử bảng mã UTF-8 khi xuất ra csv, đồng thời sử dụng tính năng  import data có sẵn trên Excel.
* Source code chỉ mới đơn giản việc crawl dữ liệu chưa thực hiện được tính năng cập nhập giá khi giá sản phẩm thay đổi theo thời gian.

**4. Dự kiến công việc tuần tới:**

* Làm sạch dữ liệu.
* Lưu trữ dataset vào trong database.
* Xây dựng cách truy vấn vào dữ liệu.
* Tìm hiểu về một số kỹ thuật phân tích dữ liệu.